Final Examination

Econ 103, Statistics for Economists

May 1, 2013

You will have 120 minutes to complete this exam. Graphing calculators, notes, and textbooks are not permitted.

I pledge that, in taking and preparing for this exam, I have abided by the University of Pennsylvania's Code of Academic Integrity. I am aware that any violations of the code will result in a failing grade for this course.

Name: _____

Student ID #: _____

Signature: _____

Question:	1	2	3	4	5	6	7	8	9	10	Total
Points:	10	20	10	25	15	20	30	10	30	30	200
Score:											

Instructions: Answer all questions in the space provided, continuing on the back of the page if you run out of space. Show your work for full credit but be aware that writing down irrelevant information will not gain you points. Be sure to sign the academic integrity statement above and to write your name and student ID number on *each page* in the space provided. Make sure that you have all pages of the exam before starting.

Warning: If you continue writing after we call time, even if this is only to fill in your name, twenty-five points will be deducted from your final score. In addition, two points will be deducted for each page on which you do not write your name and student ID.

You may use the approximation `qnorm(0.975)` ≈ 2 to simplify your calculations on this exam.

1. You are about to bid on a poster autographed by the cast of *The Jersey Shore*. This is a one-of-a-kind item, not available in stores. The most you would be willing to pay is \$500 dollars. In an ordinary auction, you could start with a low bid, say \$10, and increase it only if someone outbid you. This auction, however, is a *sealed bid auction*. Each participant submits a single bid in a sealed envelope, so there is no way for you to know how much anyone else has bid. The highest bidder wins the auction and pays whatever bid she placed in her envelope; the losers pay nothing. Based on research you have carried out on other auctions of Jersey Shore memorabilia, you estimate that the probability p that you will win the auction as a function of your bid b is as follows:

$$p(b) = \begin{cases} b/600 & b \leq 600 \\ 1 & b > 600 \end{cases}$$

   (a) (6 points) Suppose you want to maximize your *expected payoff*. Losing the auction gives you a payoff of \$0; winning gives you a payoff of \$500 *minus your bid*, i.e. your consumer surplus. How much should you bid, and what is the probability that you will win the auction? Be sure to check the second order condition.

Solution: If you win, your payoff is $500 - b$. This occurs with probability $p = b/600$. If you lose, you get zero. Hence, your expected payoff is

$$(500 - b) \cdot p + 0 \cdot (1 - p) = (500 - b) \cdot \frac{b}{600} = \frac{5b}{6} - \frac{b^2}{600}$$

We want to maximize this over b. The first order condition is

$$\begin{aligned} \frac{5}{6} - \frac{b}{300} &= 0 \\ b^* &= 250 \end{aligned}$$

Since the second derivative is $-1/300$, we know this is a maximum. With a bid of \$250, the probability of winning the auction is $250/600 \approx 0.42$.

   (b) (4 points) Using what you know about expected value, briefly explain why someone might rationally decide *not* to follow the bidding strategy you derived in part (a).

Solution: If this is really a one-of-a-kind item, it may make sense to bid more than \$250. For example, to mazimize your chance of winning the auction, you should bid \$500. The expected value bidding rule imagines a large number of *identical repeat auctions* and maximizes your *average* payoff over these auctions. If you were repeatedly bidding on a commodity (say used cars at a police auction), the expected value rule makse sense. In this case it might not. This is

similar to the St. Petersburg Paradox we studied in class. Although we calculated that the expected value of this game is infinite, no one in the class was willing to pay more then a few dollars to play the game *once*.

2. Suppose that X is a continuous random variable with probability density function

$$f(x) = \begin{cases} 3x^2 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

(a) (2 points) What is the support of X?

Solution: The support is the interval from zero to one $[0, 1]$ since the pdf is zero everywhere else.

(b) (3 points) Given your answer to (a), why *couldn't* the pdf be $4x^2$ in this example?

Solution: A pdf has to integrate to one. We have: $\int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 4x^2\,dx = 4(x^3/3)|_0^1 = 4/3$ which is *greater than one*.

(c) (6 points) Calculate the cumulative distribution function of X.

Solution: $\int_{-\infty}^{x_0} f(x)\,dx = \int_0^{x_0} 3x^2\,dx = x^3|_0^{x_0} = x_0^3$. Hence,

$$F(x_0) = \begin{cases} 0 & \text{for } x_0 < 0 \\ x_0^3 & \text{for } 0 \leq x_0 \leq 1 \\ 1 & \text{for } x_0 > 1 \end{cases}$$

(d) (3 points) Calculate $E[X]$.

Solution:

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_0^1 3x^3\,dx = \left.\frac{3x^4}{4}\right|_0^1 = 3/4$$

(e) (3 points) Calculate $E[X^2]$.

Solution:

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_0^1 3x^4\,dx = \left.\frac{3x^5}{5}\right|_0^1 = 3/5$$

(f) (3 points) Calculate $Var(X)$.

Solution:

$$Var(X) = E[X^2] - (E[X])^2 = 3/5 - 9/16 = 3/80$$

3. Let X and Y be two RVs with $Var(X) = \sigma_X^2$, $Var(Y) = \sigma_Y^2$ and $Corr(X, Y) = \rho$.

(a) (2 points) Write down an expression for $Var(X + Y)$ in terms of σ_X, σ_Y and ρ.

Solution: $Var(X+Y) = Var(X)+Var(Y)+2Cov(X,Y) = \sigma_X^2+\sigma_Y^2+2\sigma_X\sigma_Y\rho$

(b) (2 points) Write down an expression for $Var(X - Y)$ in terms of σ_X, σ_Y and ρ.

Solution: $Var(X-Y) = Var(X)+Var(Y)-2Cov(X,Y) = \sigma_X^2+\sigma_Y^2-2\sigma_X\sigma_Y\rho$

(c) (6 points) Which is larger: $Var(X + Y)$ or $Var(X - Y)$? Explain briefly.

Solution: If $\rho > 0$ then $2\sigma_X\sigma_Y\rho > 0$, hence $Var(X + Y) > Var(X - Y)$. If $\rho < 0$, then $2\rho\sigma_X\sigma_Y < 0$, hence $Var(X + Y) < Var(X - Y)$. If $\rho = 0$ then $2\rho\sigma_X\sigma_Y = 0$, hence $Var(X + Y) = Var(X - Y)$.

4. Suppose that $X_1 \sim N(\mu, \sigma^2)$ independently of $X_2 \sim N(\mu, 3\sigma^2)$. Let $\bar{X} = (X_1 + X_2)/2$.

(a) (4 points) Calculate the variance of $\bar{X}$.

Solution: In this example,

$$Var(\bar{X}) = Var\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{4}\left[Var(X_1) + Var(X_2)\right] = \frac{1}{4}\left(\sigma^2 + 3\sigma^2\right) = \sigma^2$$

(b) (4 points) Let $\tilde{\mu} = \omega X_1 + (1 - \omega)X_2$ where $\omega \in [0, 1]$. Is $\tilde{\mu}$ an unbiased estimator of μ? Prove your answer.

Solution: Yes:

$$E[\tilde{\mu}] = E[\omega X_1 + (1 - \omega)X_2] = \omega\mu + (1 - \omega)\mu = \mu$$

(c) (5 points) Define $\tilde{\mu}$ as in part (b). Calculate the variance of $\tilde{\mu}$.

Solution:

$$Var(\tilde{\mu}) = Var[\omega X_1 + (1-\omega)X_2] = \omega^2\sigma^2 + 3(1-\omega)^2\sigma^2$$

(d) (8 points) What value of ω minimizes $Var(\tilde{\mu})$? What is the minimum achievable variance? Be sure to check the second order condition.

Solution: We choose ω to minimize $\omega^2\sigma^2 + 3(1-\omega)^2\sigma^2$, yielding the FOC:

$$\begin{aligned} 2\omega\sigma^2 - 6(1-\omega)\sigma^2 &= 0 \\ 2\sigma^2\left[\omega - 3(1-\omega)\right] &= 0 \\ \omega - 3 + 3\omega &= 0 \\ 4\omega &= 3 \\ \omega &= 3/4 \end{aligned}$$

Checking the SOC:

$$2\sigma^2 + 6\sigma^2 = 8\sigma^2 > 0$$

Thus, the minimum variance is:

$$Var\left(\frac{3}{4}X_1 + \frac{1}{4}X_2\right) = \frac{9}{16}\sigma^2 + \frac{1}{16} \times 3\sigma^2 = \frac{12}{16}\sigma^2 = \frac{3}{4}\sigma^2$$

(e) (4 points) Is the sample mean an efficient estimator of μ in this example? Explain.

Solution: Although X_2 gives us information about the mean μ this information is "twice as noisy" as the information contained in X_1. Hence, by giving X_2 a lower weight than X_1, we achieve an estimator with a lower variance. In this example the sample mean is NOT efficient because there is another unbiased estimator with a lower variance, namely $3X_1/4 + X_2/4$. We know this because we showed in part (d) that

$$Var\left(\frac{3}{4}X_1 + \frac{1}{4}X_2\right) = \frac{3}{4}\sigma^2$$

whereas, from part (a)

$$Var(\bar{X}_n) = \sigma^2$$

The point here is that X_1 and X_2 are *not* identically distributed although they are independent.

5. This question asks you to supply R code to carry out a simple simulation experiment using what you learned in recitation.

   (a) (5 points) The R command `runif(n)` returns a random sample of `n` Uniform$(0, 1)$ observations. Using this information, write an R function called `unif.sim` that takes a single argument `n` and carries out the following: (i) generate a sample of `n` iid Uniform$(0, 1)$ observations, (ii) return the sample mean of these observations.

   Solution:

```
unif.sim <- function(n){
x <- runif(n)
return(mean(x))
}
```

   (b) (5 points) Suppose you wanted to run your function from part (a) 10,000 times, each with a sample size of 10, and store the result in a vector called `sims`. How could you implement this in R?

   Solution:

```
sims <- replicate(10000, unif.sim(10))
```

   (c) (5 points) Suppose you entered the command `mean(sims)` after carrying out (b). Approximately what value would you get? Why?

   Solution: This simulation examines the sampling distribution of the sample mean when the population is Uniform$(0, 1)$. We know that the sample mean is an unbiased estimator of the population mean, so its sampling distribution should be centered at 0.5, the mean of the Uniform$(0, 1)$ distribution. By the law of large numbers, the value from the simulation should be very close to this.

6. I wanted to find out how many fish are in the lake, so I sent Garth and Naijia out to catch a random sample of 100 fish, tag them, and then release them. A week later, I sent Garth and Naijia back to the lake to catch another random sample of 100 fish. Of the 100 fish they caught the second time, 20 had tags.

   (a) (5 points) Construct an approximate 95% confidence interval for the proportion of fish in the lake that have tags. Use the textbook CI to keep the calculations simple.

Solution: The textbook standard error for a sample proportion is

$$\sqrt{\widehat{p}(1-\widehat{p})/n} = \sqrt{0.2 \times 0.8/100} = 0.04$$

Here, $\widehat{p} = 20/100 = 0.2$. Hence, an approximate 95% confidence interval for p is given by 0.2 ± 0.08, in other words $(0.12, 0.28)$.

(b) (3 points) Using the fact that we know that exactly 100 fish in the lake have been tagged, what is our estimate of the *total number of fish in the lake*?

Solution: We estimated that 20% of the fish in the lake are tagged. That is,

$$\frac{\#\text{of Tagged Fish}}{\#\text{Total Number of Fish}} \approx \frac{1}{5}$$

But we know that 100 fish have been tagged:

$$\frac{100}{\#\text{Total Number of Fish}} \approx \frac{1}{5}$$

Multiplying through, we estimate that there are 500 fish in the lake.

(c) (8 points) Using your answers from above, construct an approximate 95% confidence interval for the *total number of fish in the lake*.

Solution: Let p denote the true, unknown proportion of fish in the lake that are tagged. If we knew p we could simply divide 100 by p to find the exact number of fish in the lake. If (a, b) is a 95% confidence interval for p, then

$$\begin{aligned} P(a \leq p \leq b) &= 0.95 \\ P\left(\frac{1}{b} \leq \frac{1}{p} \leq \frac{1}{a}\right) &= 0.95 \\ P\left(\frac{100}{b} \leq \frac{100}{p} \leq \frac{100}{a}\right) &= 0.95 \end{aligned}$$

Hence $(100/a, 100/b)$ is a 95% for the number of fish in the lake. Plugging in the endpoints from part (b), namely $a = 0.12$ and $b = 0.28$, our 95% confidence interval for the number of fish in the lake is approximately $(357, 833)$.

(d) (4 points) Does the CI from part (c) take the form Estimate ± ME? Explain.

Solution: No: $833 - 500 = 333$ but $500 - 357 = 143$ so our estimate is *not* in the middle of the CI. Although our CI for p was symmetric, we *transformed* the endpoints in a way that did not preserve this symmetry to construct our CI for the number of fish in the lake.

7. Let $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$ and suppose we know that $\sigma^2 = 1$.

   (a) (2 points) Write down the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)$.

   Solution: $N(0, 1)$

   (b) (6 points) Using your answer to (a), derive the sampling distribution of $\sqrt{n}\bar{X}_n$.

   Solution:

   $$\sqrt{n}(\bar{X}_n - \mu) = \sqrt{n}\bar{X}_n - \sqrt{n}\mu$$

   Hence,

   $$\sqrt{n}\bar{X}_n = \sqrt{n}(\bar{X}_n - \mu) + \sqrt{n}\mu = N(0, 1) + \sqrt{n}\mu \sim N(\sqrt{n}\mu, 1)$$

   (c) (6 points) Suppose we wanted to test the null hypothesis that $\mu = 0$ against the two-sided alternative at the 5% level. What test statistic should we use and what should our decision rule be?

   Solution: Under the null,

   $$T = \sqrt{n}\bar{X}_n \sim N(0, 1)$$

   and we reject if $|T| \geq$ `qnorm(0.975)` ≈ 2.

   (d) (6 points) How would your answer to (c) change if we tested instead against the one-sided alternative that $\mu > 0$?

   Solution: We still examine the test statistic $T = \sqrt{n}\bar{X}_n$ except in this case the decision rule is: reject if $T \geq$ `qnorm`(0.95). This critical value is *less* than 2.

   (e) (10 points) Derive the power of the hypothesis test from (c) if $\sigma = 1$, $n = 25$ and $\mu = 1$. Your answer should be given in terms of the relevant R commands.

Solution: Regardless of the true value of μ, we know from the solution to part (c) that $\sqrt{n}\bar{X}_n \sim N(\sqrt{n}\mu, 1)$. Power equals the probability of rejecting the null *when it is false*. Here we are asked to suppose that the true mean is one rather than zero, as assumed under the null, and that the sample size is 25. Hence,

$$T = \sqrt{n}\bar{X}_n \sim N(5, 1)$$

The decision rule from above was: reject if $|T| \geq 2$, so we simply need to calculate the probability that a $N(5, 1)$ random variable is greater than 2 or less than -2. Since these are mutually exclusive events, the probabilities sum. Hence, the R command to calculate the power are is follows:

```
pnorm(-2, mean = 5, sd = 1) + (1 - pnorm(2, mean = 5, sd = 1))
```

8. (10 points) A 1981 study published in the *New England Journal of Medicine* looked at the use of cigars, pipes, cigarettes, alcohol, tea, and coffee by patients with pancreatic cancer, concluding that there was "a strong association between coffee consumption and pancreatic cancer." This study was immediately criticized for carrying out multiple tests but only reporting the most statistically significant result. Subsequent studies failed to confirm an association between coffee drinking and pancreatic cancer. Suppose that six independent tests are conducted, in each case involving a product that is, in fact, unrelated to pancreatic cancer. What is the probability that at least one of these tests will find an association that is statistically significant at the 5% level?

Solution: By the complement rule, the probability is $1 - (0.95)^6 \approx 0.2649$

9. This question is based on a recent paper examining how "organic" labeling changes people's perceptions of different food products. Researchers recruited volunteers at a local mall in Ithaca, New York and gave each two samples of yogurt to taste. Although both yogurts were in fact identical, the volunteers were *told* that one of them was organic while the other was not. After tasting both, each volunteer was asked to estimate how many calories each of the samples of yogurt contained. (Since, unknown to the volunteer, both samples contained exactly the same kind of yogurt, each in fact contained the same number of calories.) To prevent confounding from anchoring or other behavioral effects, the order in which a given volunteer tasted the two yogurts, i.e. "organic" first or "organic" second, was chosen at random. The results of this experiment are stored in an R dataframe called `yogurt`. Here are the first few rows:

```
> head(yogurt)
  regular organic
1      60      40
2       5       0
3     200     100
4      60      40
5     100     100
6      90      90
```

Each row in this dataframe corresponds to a single individual's guess of the number of calories contained in each of the two yogurts. For example, the values 60 and 40 in row 1 mean that volunteer number one guessed that the regular yogurt sample contained 60 calories and the organic sample contained 40. Summary statistics for the two columns are as follows:

	`regular`	`organic`
Sample Mean	113	90
Sample Var	3600	2916
Sample SD	60	54
Sample Corr.	0.8	
Sample Size	115	

(a) (4 points) Give the units of each of the summary statistics from above:

Sample Mean _____
Sample Var. _____
Sample SD _____
Sample Corr. _____

Solution: calories, calories^2, calories, unitless.

(b) (4 points) Sara thinks that this experiment should be analyzed as independent samples data. Assume that she is correct and construct an approximate 95% CI for the difference of means (`regular` - `organic`) based on the CLT.

Solution: The difference of means (regular minus organic) is 23 calories. Sara calculates her standard error assuming independent samples:

$$\sqrt{\sigma_X^2/n + \sigma_Y^2/m} = \sqrt{3600/115 + 2916/115} = \sqrt{6516/115} \approx 7.5$$

so her confidence interval is approximately 23 ± 15, in other words $(8, 38)$.

(c) (6 points) Kevin thinks that this experiment should be analyzed as matched pairs data. Assume that he is correct and construct an approximate 95% CI for the difference of means (`regular` - `organic`) based on the CLT.

Solution: Kevin takes into account the sample correlation between columns when calculating his standard error. He does this by using the sample statistics from the table to calculate the sample variance of the *differences*: regular minus organic. In particular, he calculates:

$$s_D^2 = 3600 + 2916 - 2 \cdot 0.8 \cdot 60 \cdot 54 = 1332$$

which gives a standard error of

$$\sqrt{s_D^2/n} = \sqrt{1332/115} \approx 3.4$$

This is the only difference between his procedure and Sara's. Hence, Kevin's confidence interval is approximately 23 ± 6.8, in other words $(16.2, 29.8)$.

(d) (6 points) How do the confidence intervals constructed by Sara and Kevin differ? Explain the source of the discrepancy. Which of them has constructed the appropriate confidence interval for this example?

Solution: Kevin is right and Sara is wrong. This is matched pairs data because each row corresponds to a *single individual*. Unsurprisingly, we find a high sample correlation between the two columns: individuals who overestimate caloric content for one yogurt sample tend to do so for the other, as do individuals who underestimate. The only difference between Kevin and Sara's confidence intervals comes from how they calculated their standard errors. Both intervals are correctly centered, but Sara's is *too wide* because she calculated the standard error assuming independence between the two samples. When the sample correlation is positive this results in an *overestimate* of the standard error.

(e) (6 points) Suppose that Kevin wanted to carry out a two-sided test of the null hypothesis that organic labeling does not affect consumer's estimates of caloric content, on average. What is his test statistic? What R command should he use to calculate the p-value for his test? Will his result be greater or less than 0.05?

Solution: Kevin's test statistic is the difference of means divided by the standard error, namely $23/3.4 \approx 6.8$. To calculate the p-value in R, he should use

the command:

```
2 * (1 - pnorm(6.8))
```

The result will be less than 0.05 since the test statistic is larger than 2. Another way to see this is that his confidence interval does not include zero.

(f) (4 points) Using what you know about experiments, observational studies, hypothesis testing, and confidence intervals, what conclusions can we draw from this study?

Solution: It appears that merely labeling a product "organic" causes consumers to assume that this product contains fewer calories. Because this is a randomized experiment (randomly assigning labels to identical samples of yogurt and randomizing the order in which subjects tasted), we don't have to worry about confounding. It is less clear, however, whether this result would generalize to foods other than yogurt. Further, people from Ithaca New York who visit the mall and volunteer for a taste test may not be representative of US consumers as a whole. Ideally we would repeat this experiment using different subject pools and different foods to see how robust the result is.

10. This question refers to the four sets of regression results presented in Table 1, which appears on the final page of this exam. Each regression uses a data frame called `income.data`, the first few rows of which are as follows:

```
 head(income.data)
  income female   AFQT
1    5.5      1  6.841
2   65.0      0 99.393
3   19.0      0 47.412
4   36.0      1 44.022
5   65.0      0 59.683
6    8.0      0 72.313
```

Each row corresponds to an individual. The column `income` gives that individual's income (in thousands of dollars) in 2005. The column `female` is a dummy variable that takes the value 1 if a given individual is female. Finally, the column `AFQT` gives the individual's percentile score on the Armed Forces Qualifying Test.

(a) (3 points) Interpret the intercept in Regression 1.

Solution: The intercept in this model is the income we would predict (in thousands of dollars) for someone whose percentile score on the AFQT is zero: that is, the person who got the *minimum* recorded score on the test.

(b) (3 points) Interpret the coefficient `AFQT` in Regression 1.

Solution: Suppose we had two individuals who scores on the AFQT differed by one percentile. Then we would predict that the higher-scoring individual would earn about $520 dollars more, per year.

(c) (2 points) What is the sample correlation between income and percentile score on the AFQT?

Solution: From the output of Regression 1, $\sqrt{0.09} = 0.3$

(d) (2 points) Using the regression results, what is the average income of the men in the sample? What is the average income of the women in the sample?

Solution: From the output of Regression 2, the average income for men is $63,320, while the average income for women is $63320 - 28110 = $35,210.

(e) (5 points) Using sex *alone* as a predictor, about how accurately can we predict an individual's income? How does this compare to using AFQT score alone?

Solution: From the residual standard deviations of Regressions 1 and 2, we see that sex and AFQT percentile are roughly equivalent in their predictive power: each predicts income to an accuracy of about $45,500, i.e. neither is a particularly good predictor.

(f) (5 points) Using the regression results, construct an approximate 95% confidence interval for the difference of mean income in the population (men - women). Interpret your results.

Solution: The estimated difference (women minus men) is about -28 with a standard error of about 1.8. This gives an approximate 95% confidence interval of -28 ± 3.6, in other words $(-31.6, -24.4)$. Women earn, on average, between $31,600 and $24,400 less per year than men.

(g) (5 points) Regressions 2–4 each examine the relationship between sex and income.

How do the regression models used in each differ? You don't need to discuss the results, just the models.

Solution: Regression 2 simply compares the average incomes of men and women in the sample. Regressions 3 and 4 compare the income of men and women while holding AFQT percentile scores *fixed*. Regression 3 does this by allowing a different intercept in the relationship between income and AFQT, while Regression 4 allows both a difference intercept and a different slope.

(h) (5 points) Is there evidence of a *different* relationship between intelligence as measured by AFQT and income for men and women? If so, explain the difference.

Solution: Yes: whereas we predict an increase of \$650 per year for each additional percentile point on the AFQT for men, we predict an increase of $650 - 310 = 340$ per year for each additional percentile point for women. This difference is highly statistically significant: the standard error of the parameter that controls the difference of slopes is 0.06 while the estimate is -0.31. This corresponds to an approximate 95% confidence interval of -0.31 ± 0.12 which is bounded well away from zero.

Table 1: Regression Results

Regression 1:

```
lm(formula = income ~ AFQT)
            coef.est coef.se
(Intercept) 21.18     1.93
AFQT         0.52     0.03
---
n = 2584, k = 2
residual sd = 44.46, R-Squared = 0.09
```

Regression 2:

```
lm(formula = income ~ female)
            coef.est coef.se
(Intercept)  63.32     1.23
female      -28.11     1.75
---
n = 2584, k = 2
residual sd = 44.57, R-Squared = 0.09
```

Regression 3:

```
lm(formula = income ~ female + AFQT)
            coef.est coef.se
(Intercept)  35.55     2.04
female      -27.09     1.67
AFQT          0.50     0.03
---
n = 2584, k = 3
residual sd = 42.36, R-Squared = 0.18
```

Regression 4:

```
lm(formula = income ~ female + AFQT + female:AFQT)
            coef.est coef.se
(Intercept)  27.47     2.55
female      -10.03     3.66
AFQT          0.65     0.04
female:AFQT  -0.31     0.06
---
n = 2584, k = 4
residual sd = 42.14, R-Squared = 0.19
```